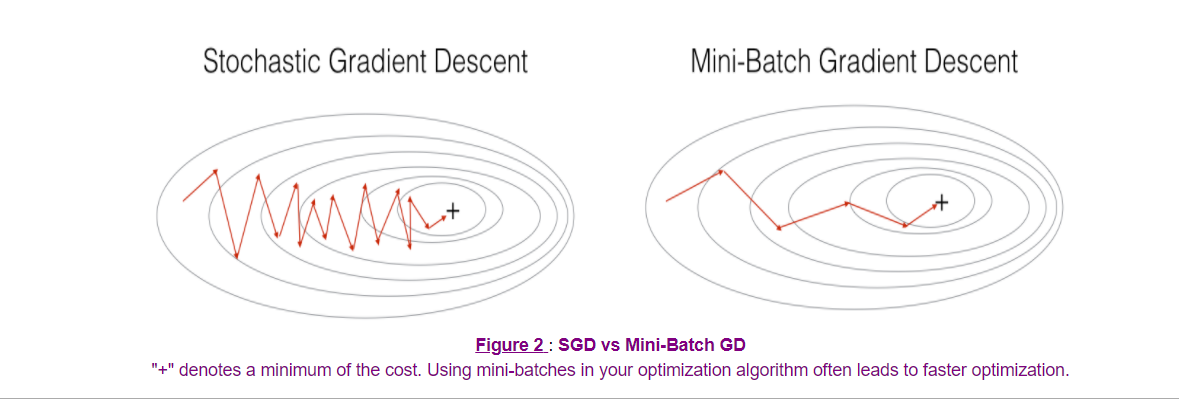
**2 - Gradient Descent**

A simple optimization method in machine learning is gradient descent (GD).





3 - Mini-Batch Gradient Descent

There are two steps:

* **Shuffle**
* **Partition**

4 - Momentum

Because mini-batch gradient descent makes a parameter update after seeing just a subset of examples, the direction of the update has some variance, and so the path taken by mini-batch gradient descent will "oscillate" toward convergence. Using momentum can reduce these oscillations.

Momentum takes into account the past gradients to smooth out the update. The 'direction' of the previous gradients is stored in the variable 𝑣 . Formally, this will be the exponentially weighted average of the gradient on previous steps. You can also think of 𝑣 as the "velocity" of a ball rolling downhill, building up speed (and momentum) according to the direction of the gradient/slope of the hill.